Использование сглаживающего функционала Тихонова для решения задачи восстановления многомерных нелинейных функций в многослойном персептроне (MLP – сетях)

А. Н. Бирюков

Башкирский государственный университет, Стерлитамакский филиал
Россия, Республика Башкортостан, 453103 г. Стерлитамак, проспект Ленина, 49.

Email: biryukov_str@mail.ru

В данной статье рассмотрен вопрос реализации концептуального базиса в конкретных методах решения обратных задач интерпретации, а также методах регуляризации с использованием принципов Байеса и Тихонова. Разработан метод оценки адекватности нейросетевых моделей в отсутствии каких-либо априорных сведений о законе распределения шумов в базе данных.

Ключевые слова: нейросетевая модель (НСМ), MLP – сети, функционал Тихонова, Байесова регуляризации, кросс-валидация.

Целью исследования является разработка методологии решения задачи аппроксимации «ядра» системы моделей бюджетирования – налогообложения регионального и муниципального уровня в нейросетевом базисе, в частности на базе многослойного персептрона (MLP).

Для простоты выходную функцию нейросетевой модели $Y(\vec{X})$, восстанавливаемую по экспериментальным данным, будем считать скалярной. Объясняющие переменные {X_j} образуют n-мерный входной вектор $\vec{X} = (x_1, x_2, ..., x_j, ..., x_n) \in R^n$, где R^n n-мерное пространство вещественных чисел. Соответственно, данные D в этом случае являются наборами примеров функциональной зависимости, т.е. парами значений $D = \{y_i, \vec{x}_i\}_{i=1}^N$,где i-номер вектор-строки наблюдения (примера); N – число наблюдений. Предполагается, что в данных присутствует шум:

$$y = h(\vec{X}, W) + \eta(\beta), \quad (1)$$

где *h (x, w)* – параметрическая многомерная функция, восстанавливающая регулярную часть скрытой в данных закономерности; *w* – совокупность параметров модели; *η(β)* – составляющая шума*; β* – параметр функции шума.

В этом случае функция $h(\vec{X}, W)$ представляет собой композицию по числу слоев сети, кроме входного слоя, операторов проецирования входных сигналов (их суммирования с синоптическими весами *w*) с последующим нелинейным преобразованием с помощью активационных функций скрытых слоев сети. Данное обстоятельство делает проблематичной возможность аналитических построений, например, представления ис-

комой функции $h(\vec{X},W)$ в виде разложения по набору базисных функций $\{\psi_k(\vec{x})\}$ решения уравнения Шредингера [3] или какого-либо другого базиса.

Что касается шумовой составляющей *η(β),* то специфика предметной области – задачи моделирования налогового контроля, а также оценки кредитоспособности предприятий – налогоплательщиков такова, что данные сильно зашумлены вплоть до сознательного их искажения. Поэтому никаких предложений о виде, распределения функции шума *η(β)* сделать нельзя. Более того, данные могут содержать противоречивые элементы (вектор – строки наблюдений). Это означает, что близким значениям вектора входа $||x_\alpha||$, $||x_\beta||$ по евклидовой норме могут соответствовать сильно различающиеся значения выхода $|y_\alpha|$, $|y_\beta|$. Это приводит к росту критерия *S* нейросетевого отображения , т.е. неустойчивости сети. Такие точки i_α, i_β должны быть отбракованы при предобработке данных.

Машинное обучение (machine learning) в нейросетях ставит своей задачей выявление закономерностей в эмпирических данных. Обучение, таким образом, относится к классу обратных задач и в общем случае является плохо определенной или *некорректной* задачей. Такие задачи отличаются особой чувствительностью некоторых решений к данным, и нахождение устойчивых решений подразумевает процедуру *регуляризации* – ограничения класса допустимых решений.

Обучающиеся модели по определению должны быть чувствительны к данным, адаптируя в процессе обучения свои настроечные параметры для наилучшего объяснения всех известных фактов. Однако хорошее качество объяснения имеющихся данных еще не гарантирует соответствующее качество предсказаний. Излишне сложные модели способны адаптироваться не только к типичным закономерностям, но и к случайным событиям в данной обучающей выборке. Как следствие, такие модели обладают плохой предсказательной способностью.

В работе [4] для задач интерпретации (восстановления) показано, что задача о минимизации стабилизатора на множестве с ограничениями типа неравенств может быть редуцирована к классической задаче на условный экстремум с ограничением вида равенств (метод Лагранжа). Такая задача, значительно более удобная для численного решения на ЭВМ, формулируется следующим образом. Пусть *λ>0* числовой параметр. Выражение:

$$J_\lambda(z) \equiv \rho^2_{\tilde{U}}(Az,\tilde{u}) + \lambda\Omega(z), z \in \overset{*}{Z}, \quad (2)$$

называется сглаживающим функционалом Тихонова для задачи интерпретации. Рассмотрим задачу:

$$z^{(\lambda)} = \arg\inf J_\lambda(z), \;\; \rho^2_{\tilde{u}}(Az^\lambda,\tilde{u}) = \delta^2, \quad (3)$$

где второе условие служит для алгоритмического выбора λ.

Задача (3), в самом деле, допускает более простой алгоритм решения. Для всякого $\lambda>0$ элемент $z^{(\lambda)}$ может быть (обычно однозначным образом) найден каким-либо прямым методом безусловной минимизации J_λ [2]; уравнение (3) есть обычное трансцендентное уравнение $\varphi(\lambda) - \delta^2 = 0$ с алгоритмически определенной левой частью, и оно может быть решено любым из известных методов на ЭВМ; определив отсюда $\lambda = \lambda(\delta)$, находим и $z_\delta = z^{(\lambda(\delta))}$. Использование указанного уравнения для определения значения параметра (λ) обычно называют методом «невязки» (невязка $\varphi(\lambda)$). Этот метод подробно изучался в работе [4].

Для замыкания задачи восстановления гиперповерхности $\Gamma(x)$ с использованием сглаживающего функционала Тихонова применительно к нейросетевому моделированию «ядра» модели следует определить δ. В качестве δ предлагается выбирать константу Липшица L как меру неоднородности данных косвенно зависящую от меры зашумления:

$$\delta \equiv L = \max_{i \in 1,N} \frac{\left\| y_2 - y_1 \right\|_{E_n}}{\left\| \vec{x}_2 - \vec{x}_1 \right\|_{E_n}}, \tag{4}$$

где $\vec{x}_2, \vec{x}_1$ – две достаточно близкие точки в базе данных, а y_1, y_2 – соответствующие им значения выхода модели:

$$\left\| \vec{x}_2 - \vec{x}_1 \right\|_{E_n} \le r, r > 0. \tag{5}$$

Как показано, некоторые обобщенно-корректные постановки обратных задач при самых общих предположениях относительно оператора A и искомого решения z и независимо от класса задачи связаны с конструкцией сглаживающего функционала Тихонова. Оказывается, что задача о минимизации этого параметрического функционала порождает целое семейство регуляризирующих операторов *(РО),* зависящих от выбора параметра λ.

В частности, для задач интерпретации (восстановления), описываемых операторным уравнением $Az=u$, существует семейство функций $\lambda = \lambda(\delta)$, таких, что экстремаль сглаживающего функционала $z^{\lambda(\delta)}$ сходится в метрике пространства Z при $\lambda \to 0$ к единственному точному решению операторного уравнения.

Если решение соответствующего операторного уравнения не единственно, то для любой из указанных зависимостей $\lambda = \lambda(\delta)$ имеет место сходимость $z^{\lambda(\delta)}$ к Ω-нормальному решению [4].

Определение. *Любой алгоритм минимизации сглаживающего функционала, построенного для обратной задачи, при заданном значении меры погрешности (или допуска) δ и каком-либо выборе зависимости $\lambda = \lambda(\delta)$, удовлетворяющей принципу регуляризации (или обеспечивающей принадлежность $z^{\lambda(\delta)}$ множеству допустимых значений), называется общим регуляризирующий оператором (РО) Тихонова.*

Для реализации общего регуляризирующего оператора Тихонова нужно решить следующие задачи:

1. Сформулировать алгоритм вычисления «прямых эффектов» т.е. невязки $\rho_{\tilde{u}}^2(Az^{\lambda},\tilde{u})$. Для обратных задач это функционал качества, вообще говоря, более общего вида, чем невязка. Отметим, что расчет прямых эффектов осуществляется многократно в ходе решения обратной задачи и занимает обычно основную долю времени ЭВМ. Ввиду этого следует позаботиться о том, чтобы соответствующий алгоритм был по возможности максимально экономичным.

2. Следует выбрать способ согласования параметра λ *с* δ: $\lambda = \lambda(\delta)$ в соответствии с принципом регуляризации.

3. Выбрать стратегию минимизации сглаживающего функционала. Отметим, что в этом отношении РО не отличается от алгоритмов, вытекающих из других корректных вариационных постановок и связан с использованием известных алгоритмов минимизации; вместе с тем структура сглаживающего функционала приводит к достаточно экономичной для ряда задач стратегии.

Один из способов согласования параметра λ с величиной δ по невязке является элементом постановки задачи (3). Поскольку задача решается на алгоритмически вводимом множестве корректности, о близости $z^{\lambda(\delta)}$ к точному решению можно судить по близости наблюдаемого и рассчитанного эффектов, что и делают при таком выборе регуляризованного приближения. Этот алгоритм является общим для задач всех типов, если для них известна величина δ.

Практический алгоритм регуляризации MLP–сетей в задаче восстановления

Предлагается следующий оригинальный алгоритм регуляризации MLP–сетей, отличающийся от известного алгоритма Harris D и Yann Le Cun [5] тем, что параметр регуляризации λ в выражении стабилизатора Тихонова оценивается на основе байесовского подхода. Ниже приводится краткое описание этого алгоритма.

В режиме обучения для MLP – нейросети решается задача восстановления многомерной параметрической функции $z(\vec{X}) \equiv \hat{Y}(\vec{X},W)$ с помощью оператора A(·), представляющего собой композицию операторов $F_3 \circ F_2$ при прямом ходе сигнала и $F_3 \circ F_4$ при обратном ходе.

Таким образом, неизвестным искомым элементом в операторном уравнении $Az = \tilde{u}$ является матрица синоптических весов нейросети *W*.

Запишем общее выражение сглаживающего функционала для искомого решения для этой обратной задачи:

$$J_{\lambda}(z) = \rho_{\tilde{u}}^2(Az,\tilde{u}) + \lambda\Omega(W), z^{(\lambda)} \in \tilde{Z}. \qquad (6)$$

Если принять в качестве оценки качества аппроксимации данных в нейросети числовую меру по:

$$\rho_{\tilde{u}}^2(Az^{\lambda},\tilde{u}) \equiv \frac{1}{2}\sum_{i=1}^{N}(\hat{y}_{p,i} - d_{p,i})^2, \tag{7}$$

а в качестве сглаживающего функционала:

$$\Omega(W) = \|z\|_{\tilde{Z}}^2 \equiv \|W\|_{E_n}^2, \tag{8}$$

то получим алгоритмически определенную вариационную задачу, решаемую в нейросетевом базисе, т.е. регуляризированный по А. Н. Тихонову алгоритм обучения:

$$w^* : J_{\lambda}(z) \equiv \left[\frac{1}{2}\sum_{i=1}^{N}\left(\hat{y}_{p.,i}(W) - d_{p,i}\right)^2 + \lambda\|W\|_{E_n}^2\right] \to \min_{W,\lambda} J_{\lambda}(z); \lambda < 1. \tag{9}$$

Здесь $||W||^2$ – обычная евклидова норма матрицы:

$$\|W\|^2 = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}^2}.$$

Параметр регуляризации λ подбирается в процессе обучения сети экспериментально.

Предлагается следующий байесовский подход для нахождения λ. Строится не одна нейросетевая модель (НСМ), а байесовский ансамбль априорных нейросетей – гипотез о порождении данных $\left\{h_q\left(\vec{X},W\right)\right\}$, различающихся архитектурой, видом активационных функций и параметром регуляризации λ. После построения ансамбля производится фильтрация НСМ внутри ансамбля по критерию ω: $q^* : P\left[(h_q(\vec{x},W,\xi^*) \mid D) \mid H\right] \geq \omega, \omega < 1$, т.е. отбраковываются сети-гипотезы, h_q, у которых отношения (N_q^*/N) числа «хорошо объясненных точек N^*» к общему числу примеров N, оказывается меньше ω. Затем в отфильтрованном ансамбле все оценки, включая оценку параметра регуляризации λ, проводятся путем осреднения по ансамблю [1].

Механизм регуляризации здесь состоит в том, что при $\lambda < 1$ и минимизации функционала $F_{\lambda}(z)$, веса уменьшаются (по сравнению с тем, если бы $\lambda = 0$). В итоге состояние S_p нейронов скрытых слоев выводятся из зоны насыщения передаточной сигмоидной функции в линейную зону. Тем самым мы усиливаем чувствительность передаточной функции нейронов $f_p(S_p)$ и, соответственно уменьшаем ее интегральные (сглаживающие) свойства (вполне непрерывного оператора. Действительно, в режиме прямого распространения сигнала (расчета) $A(\bullet) = (F_1 \circ F_2)$ на участке насыщения сигмоидной функции большим изменениям ΔSp соответствуют малые изменения выхода Δf_p (см.рис. 1).

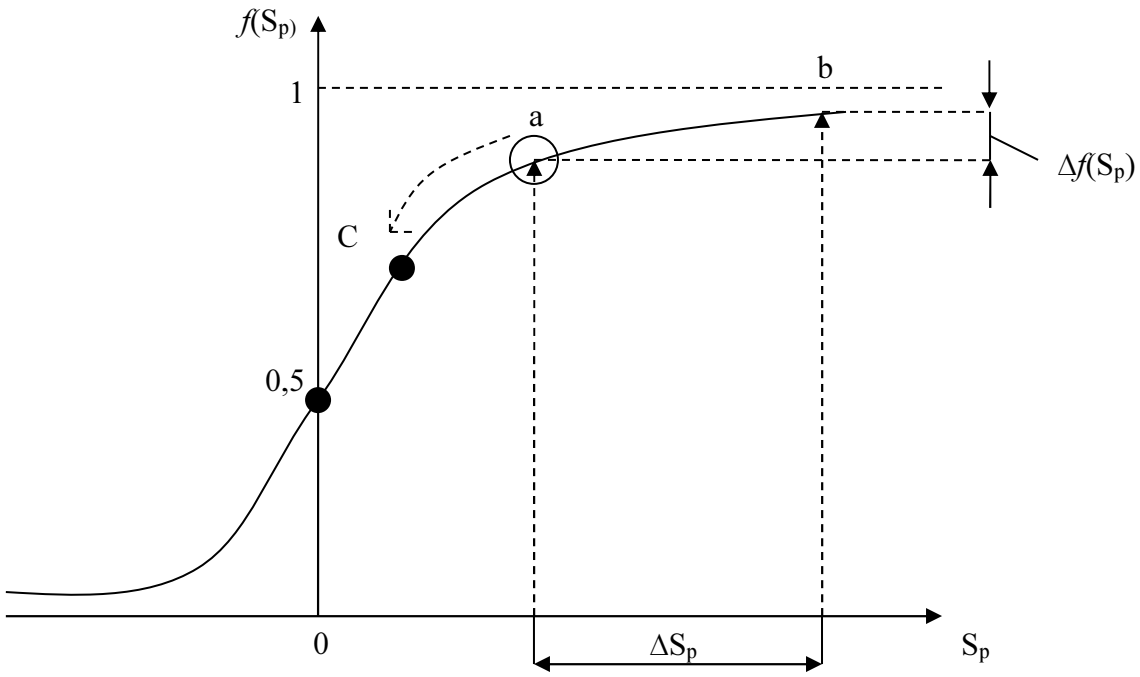

Рис. 1. Иллюстрация сглаживающих свойств передаточной функции нейрона в скрытом слое при прямом распространении сигналов.

Однако в режиме обратного распространения ошибки (модификации весов, $A(\bullet)=F_3 \circ F_4$) для того же участка сигмоидной функции (*a,b*) малым вариациям входа $\Delta f(Sp)$, например за счет возмущения данных $\{d_i\}$, будут соответствовать большие приращения функции состояния ΔSp и весов $\Delta W_{p,i}$.

Сдвиг рабочей точки сигмоидной функции из *a* в *c* на *рис.* 1 ослабляет интегральные (сглаживающие) свойства оператора алгоритма обучения и тем самым регуляризирует НСМ типа MLP-сети с алгоритмом обучения типа “обратное распространение ошибки».

На практике наибольшее распространение получили методики регуляризации, основанные на тех или иных способах оценки ожидаемой ошибки обучения на новых данных – *ошибки обобщения* E. Этот подход интуитивно кажется наиболее естественным, поскольку минимизация последней и является истинной целью обучения, тогда как практически имеем возможность измерять лишь эмпирическую *ошибку обучения*.

Такое интуитивно обоснованное обучение подразумевает два этапа: настроечные параметры модели определяются минимизацией эмпирической ошибки обучения, тогда как выбор между моделями различной сложности определяется, исходя из оценки ошибки обобщения E. Имеющиеся данные при этом также делятся на две категории.

Часть данных используют для настроек модели, а на остальных проверяют достигнутое качество обучения. Этот этап называют *валидацией* модели. Чтобы избежать зави-

симость от конкретного разбиения данных на обучающую и валидационную выборку, используют метод *кросс-валидации*, оценивая «оптимально» сложность модели в большом числе экспериментов с разными способами данных.

Трудоемкость метода кросс-валидации ограничивает его применимость, например, в системах реального времени или для действительно сложных моделей, требующих длительного обучения. В случае кросс-валидации оптимизация модели по сложности построена исключительно на эвристиках, не гарантирующих, к тому же, нахождение оптимальной модели. В байесовском подходе, напротив, количество итераций соответствует сложности данной задачи.

Байесова регуляризация служит методикой, альтернативной по отношению к кросс-валидации при оптимизации сложности модели. Она основана не на оценке ожидаемой ошибки, а на выборе наиболее *правдоподобной* модели, в пользу которой свидетельствуют имеющиеся данные. Такой подход имеет ряд преимуществ. Во-первых, он исходит из первых принципов теории вероятности и теории статистического обучения, гарантирующих уменьшение ошибки обобщения. Во-вторых, он подразумевает оценку вариаций параметров модели и соответственно – оценку точности своих предсказаний. В-третьих, поставленная таким образом задача в некоторых практически важных случаях может быть решена с минимальным числом дополнительных упрощающих предположений. И, наконец, как следствие, байесова регуляризация может быть встроена непосредственно в алгоритмы обучения. Причем, такие регуляризованные алгоритмы уже не подразумевают этапа валидации, единообразно используя все имеющиеся данные, как для выбора оптимальной сложности модели, так и для настройки ее параметров. При этом каждая итерация оптимизационного процесса улучшает модель.

Литература

1. Бирюков А. Н. Байесовский подход к регуляризации задач нейросетевого моделирования налогового и финансового подхода. Уфа: Издательство «Гилем» АН РБ, препринт №4, 2010,- с. 24.
2. Бирюков А. Н. Байесовская регуляризация нейросетевых моделей ранжирования и кластеризации экономических объектов. – Уфа: Академия наук РБ, Издательство «Гилем», 2011. – 292 с.
3. Нужный А. С., Шумский С. А. Регуляризация Байеса в задаче аппроксимации функции многих переменных// Математическое моделирование, 2003, Том 15, N: 9, с. 55–63.
4. Тихонов А. Н., Леонов А. С., Ягола А. Г. Нелинейные некорректные задачи. – М.: Наука, 1995.-312с.
5. Harris D, Yann L. C. Inprowing Generalization Performance Using Double Backpropagation //IEEE Transactions on neural Networks, 1992. Vol. 3, №6, p.p. 991–997.

Статья рекомендована к печати научной лабораторией изучения рыночной экономики экономического факультета Стерлитамакского филиала БашГУ (д.э.н., доцент, А. Н. Бирюков)

The use of Tikhonov smoothing functional for solution of the problem of reconstruction of multidimensional nonlinear functions in multilayer perceptron (MLP – networks)

A. N. Biryukov

Bashkir State University, Sterlitamak Branch
49 Lenin Street, 453103 Sterlitamak, Republic of Bashkortostan, Russia.

Email: biryukov_str@mail.ru

This article discusses the issue of implementation of conceptual basis in specific methods for solving inverse problems of interpretation and the methods of regularization using the principles of Bayesian and Tikhonov. Developed a method to assess the adequacy of neural network models in the absence of any a priori information about the distribution law of noise in the database.

Keywords: neural network model (NSM), MLP network, the Tikhonov parametric functional, Bayesian regularization, cross-validation.